Kein Nachdruck, keine Veröffentlichung im Internet oder einem Intranet ohne Genehmigung des Verlags und der GDNÄ

GDNÄ
Die Wissensgesellschaft
Gesellschaft Deutscher Naturforscher und Ärzte
German Association for the Advancement of Science and Medicine

Leopoldina
Nationale Akademie
der Wissenschaften

Vortrag Prof. Dr. Dr. Thomas Lengauer, ML, Saarbrücken
Transkript erstellt von Dr. Jörg Wetterau und
Dr. Klaus Rehfeld

Statistische Datenanalyse in der Zeit von Big Data: Leistungsfähigkeit, Risiken und Grenzen

Leopoldina Lecture

Big Data ist in aller Munde. Der Begriff ruft dabei häufig Emotionen hervor. Zum einen ist er die Basis für viel technologischen Optimismus, meist gerichtet auf neue Business-Modelle oder Versprechungen bzw. Wirkungsverstärkungen in unserem privaten und beruflichen Leben. Auf der anderen Seite ist er Grund für dystopische Perspektiven, die zum Beispiel um Durchleuchtung des Individuums und seiner Privatsphäre, übermäßige Optimierung im täglichen Leben sowie Intransparenz von Verfahren zur Entscheidungsunterstützung kreisen. Der Vortrag untersuchte die unterliegende Methodik der Datenanalyse von ihrer wissenschaftlichen Seite. Nach einer kurzen historischen Betrachtung wird aufgezeigt, was Big Data von klassischer Datenanalyse unterscheidet. Nach einer kurzen Einführung in die mathematische Methodik wird über wissenschaftliche Erfolge berichtet. Doch die Datenanalyse birgt auch Risiken, vor allem dann, wenn man die Analysen nicht sorgfältig aufstellt und ihre Resultate nicht angemessen interpretiert. Und sie hat Grenzen, vor allem, wenn es darum geht, kausale Zusammenhänge abzuleiten.

Datenanalyse gibt es nicht erst seit heute, man betreibt sie vielmehr seit vielen hundert Jahren. Tycho Brahe und sein Schüler Johannes Kepler waren frühe Datensammler, die ihre über viele Jahre gewonnenen Beobachtungsdaten über Planetenbewegungen und Sternpositionen in Tabellen – heute wären es Excel-Listen – zusammengestellt haben. Diese Datensammlung heißt „Tabulae Rudolphinae“ (1627) und kann im Internet eingesehen werden. Allein die Datenerhebung – Messungen der Planetenpositionen (bei Tycho Brahe noch ohne Fernrohr!) – war eine höchst anspruchsvolle und aufwendige Tätigkeit. Tycho Brahe starb darüber, sein Werk wurde von seinem Assistenten, Johannes Kepler, fortgesetzt. Kepler erkannte in den Daten der Planetenstellungen Gesetzmäßigkeiten und entwickelte seine berühmten Keplerschen Gesetze. Er erkannte, dass die Planetenbahnen elliptisch und nicht kreisförmig sind – das war eine völlig neue Entdeckung. Das Zentralgestirn ist dabei immer in einem Brennpunkt der Ellipse (1. Keplersches Gesetz). Das zweite Gesetz sagt auch etwas über die Umlaufgeschwindigkeit aus – in der gleichen Zeit wird eine gleichgroße Fläche überstrichen, das heißt, in großer Entfernung von der Sonne ist die Umlaufgeschwindigkeit kleiner (Abb. 1). Damit war es möglich, in die Zukunft zu schauen, und die Bewegungen der Planeten, etwa des Mars, exakt vorherzuberechnen, d.h. Kepler konnte vorhersagen, wo der Mars zu welchem Zeitpunkt steht.

„Die Datenanalyse ermöglicht also Vorhersagen, und zwar praktisch direkt. Das galt für Kepler und das gilt bis heute, etwa für alle diejenigen, die Geld mit unseren Daten verdienen, die vorhersagen, was Sie kaufen, wo sich Sie sich befinden und am Ende sogar, was Sie denken. Das ist das, was etwa Facebook macht“, brachte es Professor Lengauer auf den Punkt.

Was die Datenanalyse aber bis heute noch nicht leistet (oder nur in sehr seltenen Fällen), ist, kausale Erklärungen zu liefern. Kepler konnte beispielsweise nicht erklären, warum die Planeten sich ellipsenförmig um die Sonne bewegen. Auf diesen Sachverhalt, der bei der Interpretation von Datensätzen immer wieder zu Missverständnissen führt, verwies Lengauer in seinem Vortrag.

HINWEIS

Ein von Herrn Lengauer erstellter Artikel zum Vortragsinhalt ist erschienen unter:
Statistische Datenanalyse in der Zeit von Big Data, Nova Acta Leopoldina NF. Nr. **424**, 187 (2019). [https://www.leopoldina.org/en/publications/detailview/publication/natur-wissenschaft-gesellschaft]

©Wissenschaftliche Verlagsgesellschaft Stuttgart

Kein Nachdruck, keine Veröffentlichung im Internet oder einem Intranet ohne Genehmigung des Verlags und der GDNÄ

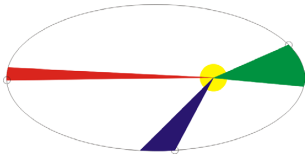
Tycho Brahe
(1546 – 1601)

Ein früher Datensammler

Tabulae Rudolphinae
(1627)

Johannes Kepler
(1571-1630)

Gesetzmäßigkeiten aus Daten



2. Keplersches Gesetz (1609)

Abb. 1. Datenanalyse in der Renaissance. Die von Tycho Brahe begonnene Erhebung von Planetenpositionen führte Kepler fort und entdeckte bei der Datenanalyse Gesetzmäßigkeiten, wie das 2. Keplersche Gesetz: Die Verbindungslinie Sonne–Planet überstreicht in gleichen Zeiträumen gleiche Flächen (Flächensatz). Ein Planet bewegt sich daher in Sonnennähe rascher als in Sonnenferne. [Abb. T. Brahe: Eduard Ender 1822-1883, Abb. Kepler: NN, Benediktinerkloster Kremsmünster, 1610 – PD, Abb. Tabulae Rudolphinae: UB Uni Kiel – Sign. Urn:hbn:de:gbr:8:2-1638239]

„Wichtig ist zu verstehen, dass es einen Unterschied zwischen Assoziation und kausalem Zusammenhang gibt. Assoziationen sind Gesetzmäßigkeiten, von der Art, wie sie Kepler gefunden hat. Den kausalen Zusammenhang kennen wir mittlerweile auch, dazu waren weitere wissenschaftliche Untersuchungen nötig.“

Entscheidende Anstöße gab Isaac Newton, der – unter anderem angeregt durch die Keplerschen Gesetze – sein berühmtes Gravitationsgesetz entwickelte. *„Aus diesem Gesetz lassen sich die Gesetze von Kepler ableiten, umgekehrt geht das jedoch nicht. Das Gravitationsgesetz hat auch eine größere Vorhersagekraft – mit ihm lässt sich z.B. auch berechnen, wie ein Projektil oder ein Flugzeug fliegt. Man sieht also, mit der Datenanalyse hört der Wissenserwerb nicht auf, sondern Datenanalyse ist der erste Schritt und dann müsste es eigentlich weitergehen. Aber die große Gefahr bei Big Data ist nun, dass das eben nicht passiert.“*

Einstein hat in der langen Kette von Wissenschaftlern nach Tycho Brahe dann wiederum Newton korrigiert, wobei Albert Einstein einer der wenigen Theoretiker war, der nicht allein auf Daten setzte, sondern vor allem auf Gedankenexperimente, um seine Theorien zu entwickeln. Er brauchte nur die Konstanz der Lichtgeschwindigkeit und die Maxwell-Gleichungen. Er gehört zu den rühmlichen Ausnahmen, die viel Wissen schaffen, indem sie über die Daten hinaus denken, so Lengauer.

Was ist anders bei Big Data?

Was ist aber nun der Unterschied zwischen dem frühen Datensammler Tycho Brahe und den heutigen Big Data? *„Brahe hat ein ganzes Leben für die Sammlung seiner Daten gebraucht. Heute hat man Sonden wie die Weltraumsonde Gaia, die den gesamten Himmel durchmustern und enorm große Volumina an Daten automatisch erheben* (vgl. Vortrag Matthias Steinmetz). *Darüber hinaus fallen Daten auch ‚ungeplant‘ an: In Facebook etwa liefern Sie die Daten mehr oder weniger freiwillig oder unbewusst. Datengenerierung ist zu einem Megaprozess geworden, das Datenaufkommen ist sehr groß und vor allem: Die Datenräume sind im Allgemeinen hochdimensional geworden. Keplers Daten waren nicht hochdimensional. Es gibt nur drei Raumdimensionen und eine Zeitdimension. Vier Dimensionen – das ist nicht hochdimensional. In der Medizin geht es um zwanzigtausend, um hunderttausend Dimensionen – das ist hochdimensional! Nicht die Datenerhebung ist heute der Engpass, sondern die Datenanalyse. Diese ist von Hand längst nicht mehr möglich. Aber durch mächtige statistische Verfahren, die in den letzten Jahrzehnten entwickelt wurden, das Data Mining, sind mit Hilfe des Computers auch komplexe Muster in den Daten auffindbar.“*

Wie funktioniert Datenanalyse?

Die Datenanalyse zielt darauf ab, bestimmte Muster in den Daten zu finden. Diese Muster werden durch Analyse von existierenden Datenmengen, den sogenannten Trainingsdaten, berechnet und dann auf in der Zukunft anfallende Daten angewandt. Das Vorgehen demonstrierte Lengauer an dem anschaulichen Beispiel des Benzinverbrauchs eines PKW in Abhängigkeit vom Gewicht und dem Baujahr des Fahrzeugs (Abb. 2a). Ausgangspunkt ist also ein zweidimensionaler (euklidischer) Datenraum, in dem jeder Datenpunkt das Baujahr (Dimension 1) und das Gewicht (Dimension 2) eines Fahrzeugs angibt. Der Verbrauch – im Beispiel angegeben in Meilen pro Gallone (mpg) – ist hier die sogenannte Ausgabe. Als Analysemethode wählte Lengauer das *überwachte statistische Lernen*. Das heißt, aus den realen Daten von rund 400 Fahrzeugen (Alter, Gewicht, Verbrauch) wird ein statistisches Modell abgeleitet. Dieses Modell kann dann genutzt werden, um (1) bei Eingabe des Baujahres und des Gewichts eines weiteren Fahrzeugs dessen Verbrauch vorherzusagen, also die Ausgabe zu schätzen (Vorhersageproblem) und um (2) eine funktionale Abhängigkeit zwischen Eingabe und Ausgabe zu finden (Inferenzproblem). Man möchte also die Vorhersage (Ausgabe) basierend auf einer Formel über funktionale Zusammenhänge ableiten (vergleichbar Keplers Erkenntnis einer Abhängigkeit von Bahngeschwindigkeit und Abstand vom Zentralgestirn).

©Wissenschaftliche Verlagsgesellschaft Stuttgart

Kein Nachdruck, keine Veröffentlichung im Internet oder einem Intranet ohne Genehmigung des Verlags und der GDNÄ

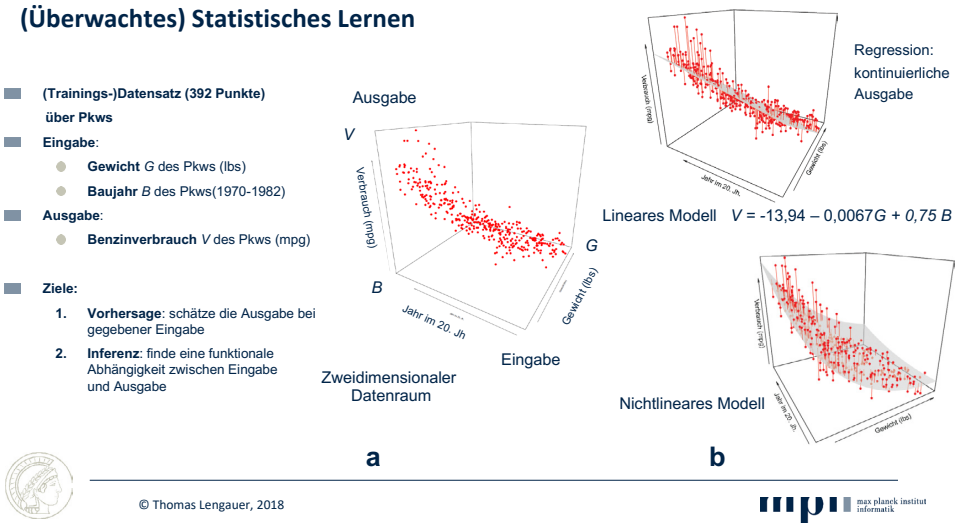


Abb. 2. Überwachtes statistisches Lernen am Beispiel des Benzinverbrauchs von Fahrzeugen. **– a.** Die sogenannten Eingabewerte Gewicht und Baujahr der Fahrzeuge liegen in einem zweidimensionalen Datenraum. Von ihnen hängt der Verbrauch ab, der in der dritten Dimension aufgetragen wird. Mit diesem Datensatz wird der Computer zunächst trainiert, um für weitere Fahrzeuge auf Grundlage von Gewicht und Alter (Eingabe) den Verbrauch (Ausgabe) vorhersagen zu können. **– b.** Lineares und nichtlineares Modell zur Vorhersage. Das nichtlineare Modell erweist sich in diesem Fall als überlegen.
[Der Datensatz stammt von T. Hastie et al.: The Elements of Statistical Learning. Springer. New York 2009]

Um einen funktionalen Zusammenhang herauszuarbeiten, wird man am einfachsten mit einem linearen Modell anfangen (Abb. 2b), in dem die Ausgabe linear von der Eingabe abhängt. Dies führt zu einer Vorhersageebene (Regressionsebene), die so gelegt wird, dass sie mitten durch die Punktwolke geht, so dass die Vorhersagefehler – repräsentiert durch die Länge der senkrechten roten Linien in der Abbildung – möglichst gering sind (genauer: man minimiert die Summe der Fehlerquadrate). Dennoch ist deutlich, dass die Vorhersagefehler je nach Eingabegröße sehr unterschiedlich sind. Auffällig ist, dass es Bereiche gibt, in denen das lineare Modell die Reichweite pro Gallone systematisch unterschätzt (bei Autos jüngeren Baujahres) bzw. überschätzt (bei Fahrzeugen mittleren Baujahres). Dies motiviert, ein nichtlineares Modell zugrunde zu legen, das den Fehler verringert und deshalb vorzuziehen ist.

Doch wie weit kann und soll man die Fehler verringern? Dass man hier nicht übertreiben darf, ist im Vorhinein einleuchtend, wenn man bedenkt, dass der Benzinverbrauch sicherlich nicht allein von den beiden Parametern Baujahr und Gewicht abhängen wird. Das Problem des „Übertrainierens“ machte Lengauer an einer anderen Modellierung deutlich, die nicht kontinuierliche, sondern klassifizierende Ausgabewerte vorhersagt (Abb. 3). Die Klassifikation stellt eine andere, wichtige Form der Datenanalyse dar, die ihre spezifischen Stärken und Probleme hat. Im Beispiel wurden nur zwei Klassen gebildet, und einfach eine Grenze zwischen sparsamen und spritfressenden Autos eingeführt, so dass die Werte nun oberhalb oder unterhalb einer Ebene liegen (Abb. 3a). In einer geeigneten Darstellung, bei der man „von oben“ auf die beiden Ebenen schaut, die Werte oberhalb (sparsam) in blau, die unterhalb (spritfressend) in rot markiert sind, zeigt sich ein linearer Zusammenhang, der eine gute Beschreibung mit nur geringen fehlerhaften Zuweisungen darstellt (Abb. 3b). Mathematisch kann man den Trainingsfehler allerdings minimieren, in diesem Fall sogar auf Null bringen. Ein Modell, das fast keinen Trainingsfehler aufweist, ist auf der Trainingsmenge somit „fehleroptimiert“, doch ist das Ergebnis unsinnig – der Benzinverbrauch *kann nicht* in dieser Weise von Gewicht und Baujahr abhängen. Man hat hier das Rauschen mittrainiert, und nicht von den tatsächlichen Signalen getrennt. Deshalb wird auch die Vorhersagekraft eines solchen übertrainierten Modells auf zukünftigen Daten schlechter sein als die des einfacheren linearen Modells. Die Kunst besteht also darin,

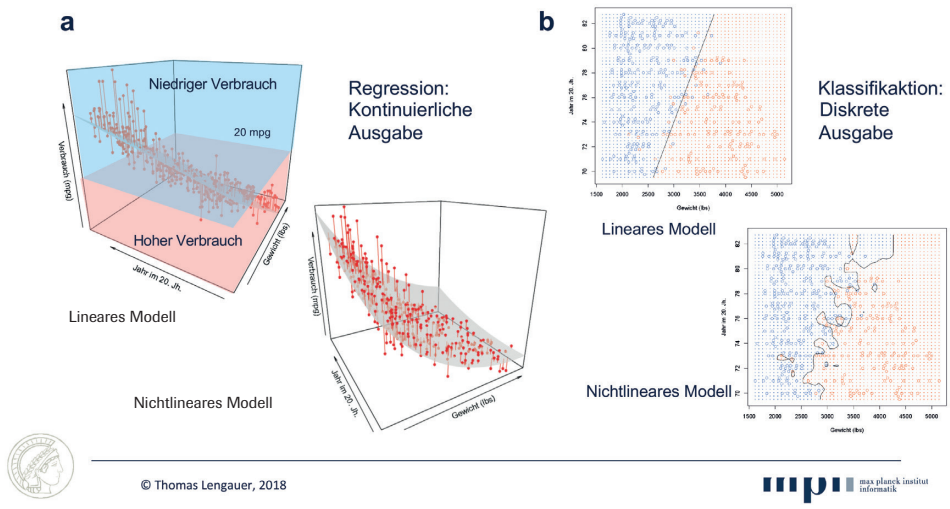


Abb. 3. Kontinuierliche und diskrete Ausgabe. **– a.** Alternativ zu Regressionsanalysen mit dem Ziel kontinuierlicher Ausgabewerte kann man die Daten auch so aufbereiten, dass klassifizierende Aussagen möglich sind. Hier geht es nur um die Aufteilung in Fahrzeuge mit niedrigem und hohen Verbrauch. **– b.** Das lineare Modell erweist sich hier als eindeutig überlegen, während das nichtlineare Modell, das mit den Trainingsdaten optimiert wurde, untauglich ist. Es hat die „Ausreißer“ mittrainiert.

©Wissenschaftliche Verlagsgesellschaft Stuttgart

Kein Nachdruck, keine Veröffentlichung im Internet oder einem Intranet ohne Genehmigung des Verlags und der GDNÄ

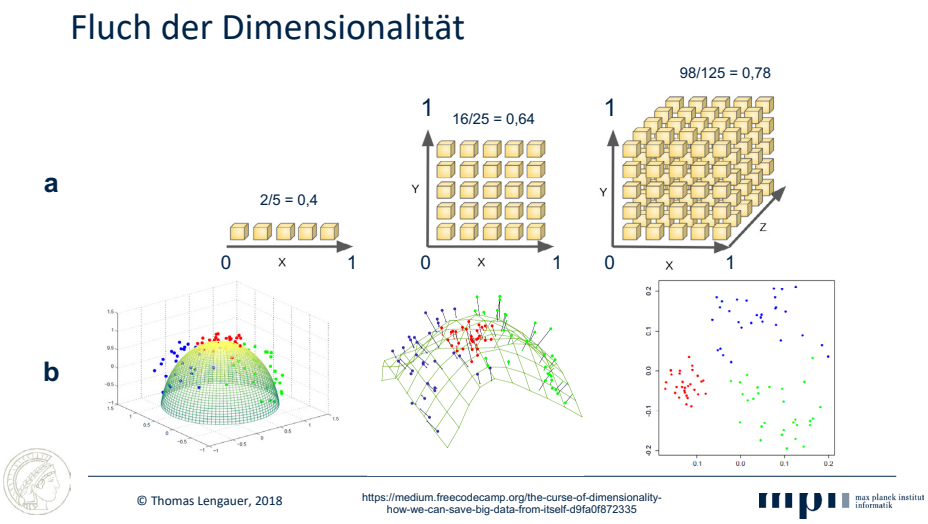


Abb. 4. Das Problem der vielen Dimensionen. **– a.** Mit steigender Dimensionszahl (1) erfordert eine konstante Datendichte (exponentiell) steigende Zahlen von Datenpunkten, (2) nimmt bei konstanter Zahl der Datenpunkte die Datendichte (exponentiell) ab, (3) befindet sich ein immer größerer Teil der Datenpunkte am Rand des Datenbereichs. **– b.** Die Dimensionsreduktion ist daher ein wesentliches Element der Datenanalyse. In dem hier gezeigten Beispiel gelingt es, die Datenwolke auf eine zweidimensionale Fläche zu projizieren. [Abb. b: T. Hastie et al.: The Elements of Statistical Learning Springer. 2009]

das Rauschen vom Signal zu trennen, oder – mit den drastischen Worten des Nobelpreisträgers Klaus von Klitzing: *„Ein Wissenschaftler muss herausfinden, was Dreck ist und was gut ist.“*

Der Fluch der Dimensionalität

Bei der Datenanalyse sind zweidimensionale und dreidimensionale Räume noch halbwegs übersichtlich und erfassbar. Hochdimensionale Datenräume sind dagegen sehr schwer zu analysieren (Abb. 4). *„Denn mit steigender Dimension erfordert eine konstante Datendichte exponentiell steigende Zahlen von Datenpunkten. Für hochdimensionale Räume werden daher unfassbar große Datenmengen benötigt, um diese Räume zu befüllen. Und wir dürfen einen zweiten Punkt nicht außer Acht lassen, das ist die geometrische Form der hochdimensionalen Räume. Das Problem dabei ist, dass mit steigender Dimension sich ein immer größerer Teil der Datenpunkte am Rand des Datenbereichs befindet. Bei einer Linie – hier gibt es nur eine Dimension – sind das z.B. bei fünf Punkten zwei am Rand, also 40%, bei einem Quadrat (zwei Dimensionen) schon 64% , bei einem Würfel (drei Dimensionen) befinden sich schon 78% am Rand und bei hochdimensionalen Körpern befindet sich irgendwann praktisch alles am Rand, aber nichts mehr in der Mitte. Das ist äußerst unvorteilhaft. Viele Methoden zur Datenanalyse schließen zum Beispiel von den Werten in der Nachbarschaft eines Datenpunktes auf den Wert des Datenpunktes (Nearest-Neighbor Methoden). Solche Nachbarschaften sind in hochdimensionalen Räumen aber bei praktisch allen Punkten nur unvollständig ausgebildet, da die Punkte am Rand des Datenbereichs liegen. Daher ist Dimensionsreduktion ein wesentliches Element der Datenanalyse. Zum Glück zeigt sich, dass viele Datenwolken, die sich in hochdimensionalen Räumen befinden, gar nicht hochdimensional sind, sondern, dass sich die Punktwolken auf wenige Dimensionen zurückführen lassen – sind dies nur zwei Dimensionen, so kann man quasi die Punkte mit einem Tuch einfangen, das man geeignet gekrümmt in der hochdimensionalen Datenraum legt, um die Datenpunkte darauf zu projizieren. Wenn man das Tuch danach glatt ausbreitet, liegen alle Punkte in einer Ebene. Damit befinden sich die Punkte in einem zweidimensionalen Raum, der gut bevölkert ist. Dimensionsreduktion ist ein wesentlicher Teil im statistischen Lernen!“*

Was Datenanalyse kann: HIV Therapie

An einem medizinischen Beispiel zeigte Lengauer, wie erfolgreich statistisch gelernte Datenanalyse sein kann. Aus seiner

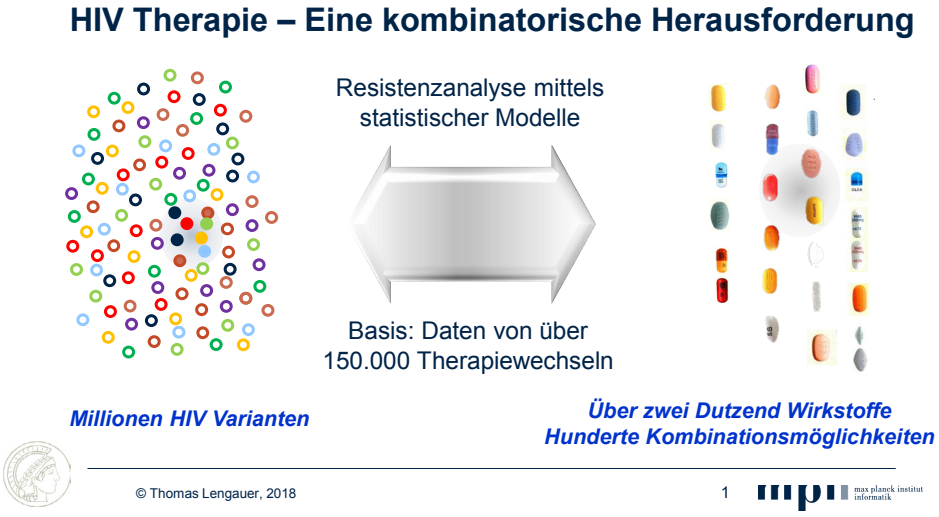


Abb. 5. Problem HIV-Infektion. Welche Medikamentenkombination sind bei welchen HIV-Varianten optimal? In der Regel werden 3 bis 4 Wirkstoffe in Kombination verabreicht, rund 30 Präparate stehen zur Verfügung.

©Wissenschaftliche Verlagsgesellschaft Stuttgart

Kein Nachdruck, keine Veröffentlichung im Internet oder einem Intranet ohne Genehmigung des Verlags und der GDNÄ

Sicht gibt es derzeit kein besseres Beispiel für die durch bioinformatische Datenanalyse unterstützte personalisierte Medizin als die Behandlung von HIV-Infektionen. Jeder Patient hat in seinem Körper unterschiedliche HIV-Viren – ihre Genome können sich bis zu 20% voneinander unterscheiden –, die wiederum mit einem gezielten Tablettencocktail mit Hunderten Kombinationsmöglichkeiten von Wirkstoffen behandelt werden (Abb. 5). Wie gewährleistet man eine auf den jeweiligen Patienten optimal abgestimmte Therapie? Hierfür hat das Max-Planck-Institut für Informatik in Zusammenarbeit mit weiteren Forschungsinstituten ein auf Datenanalyse beruhendes Konzept erarbeitet, das völlig neue Behandlungsmöglichkeiten eröffnet.

Die Patienten werden gewissermaßen mit einem Rechner eingestellt, der die bestmögliche Behandlung anhand gesammelter Daten bestimmt. Grundlage hierfür sind die Daten, die bei über 150 000 Therapiewechseln gewonnen wurden. Die Eingabe ist hochdimensional, was vor allem auf das Virus zurückgeht, das bei einer Länge von 10 000 Basenpaaren zahlreiche Mutationen aufweisen kann, die seine Wirkstoffresistenz beeinflussen. Als Ausgabe sind folgende drei Werte von besonderem Interesse:

(1) das Resistenzniveau des Virus (Laborwert, kontinuierlicher Wert),
(2) der Therapieerfolg (klinisch erhoben, binär klassifiziert) und
(3) die Dauer der Wirksamkeit der Therapie (klinisch erhoben, ein kontinuierlicher Wert).

Verschiedene Varianten der Resistenzanalyse sind auf dem Server geno2pheno.org frei verfügbar.

Nach Eingabe des Virusgenoms erhält man eine Ausgabe mit Informationen über das Resistenzniveau des jeweiligen Virus bezüglich der verfügbaren antiviralen Medikamente. Im gezeigten Beispiel für einen „austherapierten“ Patienten (Abb. 6) lagen alle Werte über vier, ein Wert der eigentlich möglichst unterschritten werden sollte, um Aussicht auf Erfolg zu haben. Nur zwei Präparate liegen zumindest in der Nähe dieses Wertes. Doch die Ausgabe enthält auch Angaben über Mutationen – und zwar über solche, die eine Resistenz gegenüber dem jeweiligen Wirkstoff verleihen (rot) und solchen, die die Resistenz gegen den Wirkstoff reduzieren (grün). Aufgrund der Daten wurde im vorliegenden Fall eine ungewöhnliche Wirkstoffkombination empfohlen: der Wirkstoff SQV – mit einem Resistenzwert (z-Score) nahe dem erforderlichen Wert 4, was naheliegend ist – und der Wirkstoff LPV (Abb. 6). Ausschlaggebend für die Wahl dieses pharmakologisch völlig unwirksamen Wirkstoffs ist ein evolutionäres Argument: LPV ist in der Therapie, um den Selektionsdruck auf die Resistenzmutation 76V (rot in der Zeile zu LPV) aufrecht zu erhalten, die die Wirksamkeit des Wirkstoffs SQV (grün in der Zeile zu SQV) erhöht. Eine solche unorthodoxe – und, wie sich herausstellte, erfolgreiche – Empfehlung hätte ohne Data Mining und eine interpretierte Aussage nicht ausgesprochen werden können!

„Daher ist die Erklärbarkeit von Vorhersagen auf der Basis von statistischen Modellen ein wichtiger, aber leider in der Presse, aber auch in der Wissenschaft viel zu wenig beachteter Aspekt der Datenanalyse. Es gibt viel zu wenige erklärbare statistische Modelle und Algorithmen. Algorithmen müssen sich erklären können. Die ‚Methode Vertrauen‘ reicht hier meiner Meinung nach nicht“, betonte er.

„Der Server ist ein großer Erfolg – doch letztlich befinden wir uns mit dieser Datenanalyse zur Behandlung von HIV auf einer ähnlichen Ebene wie Kepler. Wir wissen nicht, warum einige der Viren resistent sind und andere nicht. Wir wissen auch nicht wirklich, warum ein Wirkstoff ein HIV-Virus lahmlegt. Wir haben nur Muster gefunden darüber, wie sich Viren in der Vergangenheit verhalten haben und projizieren lediglich aus den gesammelten Daten auf eine mögliche erfolgreiche Therapie, wagen also mit dem statistischen Modell eine Vorhersage“, schränkte er ein und präzisierte: *„Die Datenanalyse von geno2pheno basiert auf Assoziationen in den klinischen Daten über beobachtete Therapieverläufe. Die molekularen Ursachen sind nicht Gegenstand unserer Analyse. Prospektive klinische Studien sind schon aussagekräftiger als Assoziationen. In ihnen werden die zu untersuchenden Prozesse nicht nur beobachtet, sondern es wird kontrolliert in sie eingegriffen. Dadurch können sie Kausalzusammenhänge aufdecken. Dennoch ist festzuhalten: Wir können erfolgreich therapieren, auch wenn wir kein vollständiges Verständnis von den unterliegenden Vorgängen haben. Das widerspricht in gewisser Weise dem Diktum von Max Planck, das zum Leitbild der Max-Planck-Gesellschaft wurde: ‚Das Verstehen muss dem Anwenden vorausgehen‘ – Aber hier haben wir einen Fall, wo wir gut daran tun, die Vorhersagekraft der statistischen Analyse zu verwenden, obwohl wir die molekularen Zusammenhänge nicht vollständig verstehen.“*

Geno2pheno Server

"rote" Mutationen erhöhen Resistenz

"grüne" Mutationen reduzieren Resistenz

Drug	RF(*)	z-score	Scored Mutations(**)
ZDV	257.276	9.945	215Y 210W 41L
ddI	4.057	6.087	184V 121H 178L 215Y 177E
d4T	2.477	4.594	215Y 178L 118I 184V 210W 121H 41L
3TC	149.029	18.504	184V 41L 215Y
ABC	7.501	11.968	184V 215Y 210W 41L
TDF	4.613	6.684	215Y 41L 98G 184V 178L 118I 177E 135T
NVP	149.235	5.243	103N 135T 210W 98G 211K
EFV	60.034	7.450	103N 98G 135T 210W 214F 177E
SQV	3.908	4.658	63P 46I 37N 71V 72T 60E 76V 57K
IDV	52.999	11.171	46I 76V 63P 82A 61E 62V 60E 71V
NFV	33.156	8.246	46I 63P 10F 3I 76V 60E 62V
APV	59.452	13.085	76V 54M 10F 46I 63P
LPV	100.108	14.521	46I 76V 10F 63P 82A
ATV	19.252	7.838	82A 46L 54M 76V 93L 71V 62V

www.geno2server.org

Abb. 6. Die Daten zur HIV-Therapie wurden so aufbereitet, dass sie Zusatzinformationen über Resistenzverleihende und -mindernde Mutationen enthalten. Nur so wurden interpretierende Vorhersagen möglich, die dabei halfen, eine unorthodoxe, aber erfolgreiche Therapie zu wählen.

©Wissenschaftliche Verlagsgesellschaft Stuttgart

Kein Nachdruck, keine Veröffentlichung im Internet oder einem Intranet ohne Genehmigung des Verlags und der GDNÄ

VOM RICHTIGEN UMGANG MIT DEN DATEN

- Datenanalyse ist eine mächtige Methode zum Auffinden von Mustern in Datenmengen jeder Größe.
- Mit Datenanalyse (Data Mining, Big Data) kann man systematisch vielversprechende Hypothesen generieren – quasi als Alternative zum genialen Einfall eines Wissenschaftlers.
- Gute statistische Modelle lernen nur das Signal in den Daten, aber nicht das Rauschen.
- Hochdimensionale Datenräume stellen besondere Anforderungen und benötigen eine sinnvolle Reduktion der Dimension.
- Durch Datenanalyse aufgedeckte statistische Assoziationen geben für sich noch keinen Aufschluss über kausale oder mechanische Zusammenhänge.

Professor Lengauer wirbt für Ursachenforschung auch im Umfeld von Big Data. *„Ich höre dann Aussagen wie: ‚Wenn in den Daten alle für eine Vorhersage relevanten Assoziationen enthalten sind und wir über Datenanalyse-Verfahren verfügen, diese zu extrahieren, warum brauchen wir dann noch Ursachenforschung und Theorie?‘ Das halte ich für eine gefährliche naive Frage und Feststellung, und praktisch die gesamte Datenanalyse auf dem Internet nimmt auf ursächliche Zusammenhänge keinerlei Bezug. Aber Wissenserwerb ist mehr als Datenanalyse“,* mahnte er an.

Assoziation vs. Kausalität

Wie leicht assoziative Datenanalysen zu falschen Annahmen von Kausalzusammenhängen führen können, ist insbesondere bei Meldungen aus dem medizinischen Bereich zu beobachten.

„Keiner von uns ist dagegen gewappnet: Die Aussagen ‚Eine Variante eines Gens X ist assoziiert mit Auftreten der Krankheit Y‘ werden wir intuitiv in der Weise zur Aussage verknüpfen: ‚X verursacht Y.‘“

Ob die Genvariante Ursache der Krankheit ist, ist mit einer assoziativen Analyse nicht gesichert.

Welche Brisanz in Aussagen zu Assoziationen stecken kann, zeigt sich bei der (in gewissen Fällen statistisch gesicherten) Aussage, dass es in der Nähe eines Kernkraftwerks gehäuft Krebsfälle gibt. Naheliegend ist der Verdacht, dass das Kernkraftwerk etwa durch Freigabe radioaktiver Strahlung die Ursache der Krebserkrankungen ist, doch erwiesen ist dies keineswegs. *„Wichtig ist, dass wir uns klar machen, dass hier nur reine Assoziationen vorliegen, und auch andere Faktoren in Erwägung zu ziehen sind, ein ‚Störfaktor‘ zum Beispiel. Im Falle von Kernkraftwerken kann es etwa das soziale Profil der Anwohner und deren damit verbundener Lebenswandel sein, die zu gehäuften Krebsfällen führen, aber auch das ist nur eine Vermutung. Aufklärung kann hier nur eine Analyse schaffen, die alle Störfaktoren einbezieht und kausale Zusammenhänge aufdeckt. Eine weniger emotionsgeladene und politische Assoziation ist die zwischen Geburtenrate und Vorkommen von Störchen. In diesem Fall könnte die Urbanisierung als Störfaktor fungieren, der sowohl die Geburtenrate als auch die Anzahl der Störche senkt.“*

Es bleibt festzuhalten:

- Die Datenanalyse allein liefert keine kausalen Erklärungen.
- Bei geeignetem Design liefert sie – wie im Fall der HIV-Therapie – informative Vorhersagen, die uns eine *Interpretation von Assoziationen* ermöglichen.
- Ein Kausalzusammenhang zwischen Eingabe und Ausgabe lässt sich in aller Regel nur durch kontrollierte Experimente, wie etwa – in der Medizin – durch klinische Studien aufdecken.
- Wir brauchen Theoriebildung, um
  - Methoden zur Ableitung kausaler Zusammenhänge aus statistischen Assoziationen zu entwickeln,
  - zu einem tiefen Verständnis zu gelangen, das weit über die Möglichkeit der Vorhersage hinausgeht (z. B. Newtons Gravitationsgesetz und Keplersche Gesetze) und
  - zu einem mechanistischen Verständnis zu kommen (in der Medizin bis zur molekularen Ebene, um an den Ursachen ansetzende Therapien zu entwickeln).

[Photo Rieken für Leopoldina]

Prof. Dr. **Thomas Lengauer** (Jahrgang 1952) studierte bis 1975 Mathematik an der FU Berlin und wurde dort 1976 promoviert. Nach einem Studium der Informatik an der Stanford University, das er ebenfalls mit einer Promotion abschloss, war er von 1979 bis 1981 bei den Bell Labs beschäftigt. Nach der Habilitation an der Universität des Saarlandes (1984) folgte er einem Ruf an die Universität Paderborn und war dann von 1992 bis 2001 Professor an der Universität Bonn, wo er auch Direktor des Instituts für Algorithmisches und Wissenschaftliches Rechnen an der GMD – Forschungszentrum Informationstechnik in Sankt Augustin war. Von 2001 bis 2018 war er Direktor des Max-Planck-Instituts für Informatik und er ist derzeit Emeritus an diesem Institut. Seit 2015 ist Lengauer Präsidiumsmitglied der Deutschen Akademie der Naturforscher Leopoldina. Gegenwärtiger Forschungsschwerpunkt ist die Bioinformatik. Er ist derzeit der Präsident der International Society for Computational Biology.

Max-Planck-Institut für Informatik, Research Group Computational Biology, Campus E1 4, 66123 Saarbrücken

©Wissenschaftliche Verlagsgesellschaft Stuttgart